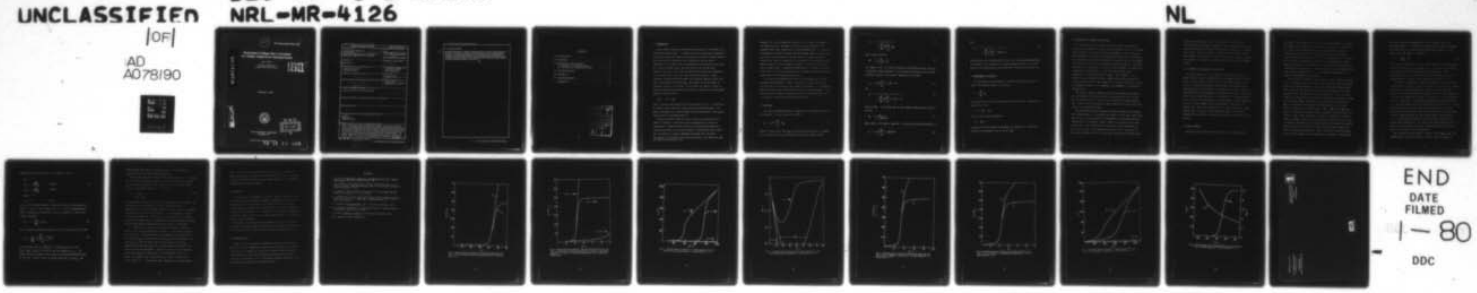
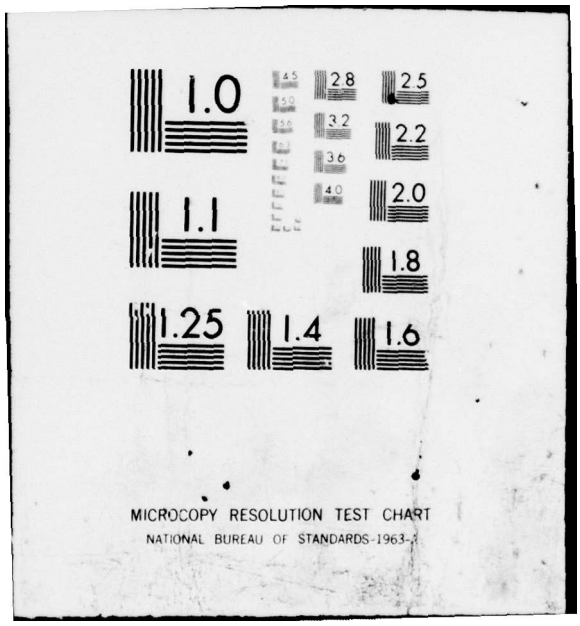
AD-A078 190 NAVAL RESEARCH LAB WASHINGTON DC F/G 12/1
PERFORMANCE COLLAPSE DUE TO OVERHEAD IN A SIMPLE, SINGLE-SERVER--ETC(U)
DEC 79 J E SHORE
UNCLASSIFIED NRL-MR-4126 NL



END
DATE FILMED
1-80
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

NRL Memorandum Report 4126

Performance Collapse Due to Overhead in a Simple, Single-Server Queuing System

JOHN E. SHORE

Information Systems Staff
Communications Sciences Division

LEVEL II

December 11, 1979

NAVAL RESEARCH LABORATORY
Washington, D.C.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NRL Memorandum Report 4126
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER: (9)
4. TITLE (and Subtitle): (6) PERFORMANCE COLLAPSE DUE TO OVERHEAD IN A SIMPLE, SINGLE-SERVER QUEUING SYSTEM
5. TYPE OF REPORT & PERIOD COVERED: Final report on an NRL problem. (13)
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): (10) John E. Shore
8. CONTRACT OR GRANT NUMBER(s): (12) 941
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Research Laboratory, Washington, DC 20375
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NRL Problem 75B02-35, 61153N, RR014-09-41
11. CONTROLLING OFFICE NAME AND ADDRESS: (11) 11 Dec 79
12. REPORT DATE: December 11, 1979
13. NUMBER OF PAGES: 23
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): (16) RR014 09; (17) RR014 09 41
15. SECURITY CLASS. (of this report): UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited. (14) NRL-MR-4126
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES:
19. KEY WORDS (Continue on reverse side if necessary and identify by block number):
Overhead
Queuing theory
Performance collapse
20. ABSTRACT (Continue on reverse side if necessary and identify by block number):
We consider a simple model of overhead in batch computer systems and message switching systems: an exponential, single-server queuing system with finite storage capacity, constant arrival rate, and queue-length-dependent service time. We consider cases in which the expected service time time consists of a constant plus a term that grows linearly or logarithmically with the queue length. We show that the performance of this system --- as characterized by the expected number of customers in the system, the expected time in the system, and the rate of missed customers --- can undergo a sudden collapse as the result of small changes in the arrival rate, the overhead rate, or the queue capacity. The system has the interesting property that increasing the queue capacity (Continues)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

251 950

20. Abstract (Continued)

cont.

can decrease performance. In addition to equilibrium results, we consider the dynamic behavior of the model. We show that the system tends to operate in either of two quasi-stable modes of operation --- one with low queue lengths and one with high queue lengths. System behavior is characterized by long periods of operation in both modes with abrupt transitions between them. We point out that the performance of a saturated system may be improved by dynamic operating procedures that return the system to the low mode.

CONTENTS

1.0 INTRODUCTION 1

2.0 THE MODEL 2

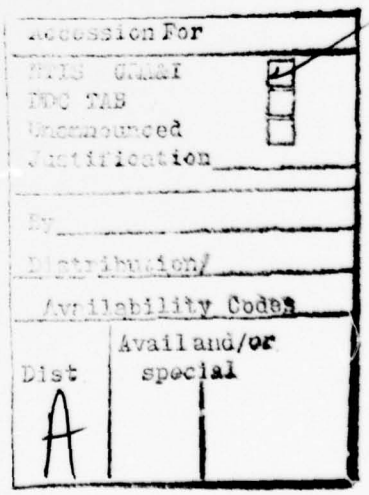
3.0 PERFORMANCE OF THE MODEL 4
3.1 Performance for Linear Service Times 5
3.2 Performance for Logarithmic Service Times 6

4.0 DYNAMIC BEHAVIOR 6

5.0 DISCUSSION 11

6.0 ACKNOWLEDGMENTS 11

REFERENCES 12



1.0 INTRODUCTION

Batch computer systems and message switching systems can be modeled [1] by the M/M/1/K queuing system --- customers arrive with independent, exponentially distributed interarrival times at an average rate λ from an infinite customer pool; they wait in a queue with finite capacity K; they are served independently by a single server with exponentially distributed service times at a constant average rate μ; and they return to the customer pool. Customers arriving at a full queue are denied service and return immediately to the customer pool. As a first step in modeling overhead, it seems reasonable to modify the M/M/1/K model from a constant expected service time $1/\mu$ to an expected service time $1/\mu_n$ that depends on the the number of customers currently in the system. In this paper, we consider a model in which the expected service time has the form of a constant plus a queue-length dependent "overhead" term,

$$1/\mu_n = f(n) + 1/\mu, \tag{1}$$

where f increases monotonically with n and satisfies f(1) = 0. In particular, we consider cases in which f(n) grows linearly and logarithmically. This model is an example of a class of models that has been studied in more general terms by Courtois and Vantilborgh [2].

Modeling overhead by means of a service time that increases with the number of customers in the system could be appropriate in a variety of circumstances. Examples include systems with scheduling algorithms that consider attributes of all waiting customers in deciding which one to serve next (as opposed to customer independent algorithms like first-come, first-served), and systems in which overhead functions like compaction, page

Note: Manuscript submitted October 5, 1979.

swapping, etc., are performed more frequently as the number of customers in the system increases. The model can also be seen as an abstraction of broadcast packet radio systems such as slotted ALOHA [3],[4]: the queue size corresponds to the number of nodes that have messages ready to transmit, and the service time corresponds to the time taken for a single successful transmission. Owing to contention, the transmission time increases with the number of nodes that have messages to transmit. In a variety of ways, the model's dynamic behavior resembles the known behavior of systems like ALOHA [3],[4], and it may be that such systems can be modeled effectively using the simple model and techniques presented herein.

It is obvious that modifying the service rate from $\mu_n = \mu$ to (1) will decrease system performance. What is surprising, at least to me, is how non-linear the relationship is: Huge changes in performance --- as measured by mean queue length, mean waiting time, and rate of rejected customers --- can result from small changes in the arrival rate λ, in the queue capacity K, or in a parameter characterizing the derivative df(n)/dn.

2.0 THE MODEL

Let p_n be the probability that n customers are in the system (waiting or being served). Then p_n is given [5, p. 92] by

$$p_n = p_o \lambda^n \prod_{m=1}^{n} 1/\mu_m \quad , \tag{2}$$

where λ is the arrival rate, μ_n is the service rate given that n customers are in the system, and p_o is the probability that the system is empty,

$$p_o = \frac{1}{1 + \sum_{n=1}^{K} \lambda^n \prod_{m=1}^{n} 1/\mu_m} \quad . \tag{3}$$

Now, let μ_n be given by

$$\mu_n = \frac{\mu}{(1 + \alpha(n - 1))} \quad . \tag{4}$$

For example, if $\alpha = .01$, then (4) states that the expected service time $1/\mu_n$ is $1/\mu$ when a single customer is in the system and increases by 1% for each additional customer in the system. Combining (2)-(4) yields

$$p_n = p_o \rho^n \prod_{m=1}^{n} (1 + \alpha(m - 1)) \tag{5}$$

and

$$p_o = \frac{1}{1 + \sum_{n=1}^{K} \rho^n \prod_{m=1}^{n} (1 + \alpha(m - 1))} \quad . \tag{6}$$

where $\rho = \lambda/\mu$. As a second case, we also consider expected service times of the form

$$\mu_n = \frac{\mu}{(1 + \beta \log(n))} \quad , \tag{7}$$

where log(n) is the natural logarithm. In this case the queue distribution is

$$p_n = p_o \rho^n \prod_{m=1}^{n} (1 + \beta \log(m)) \tag{8}$$

with

$$p_o = \frac{1}{1 + \sum_{n=1}^{K} \rho^n \prod_{m=1}^{n} (1 + \beta \log(m))} \quad . \tag{9}$$

The motivation for considering both (4) and (7) is that algorithms depending on attributes of all customers in the queue probably execute in times that grow at least logarithmically and no more than linearly with the queue size.

3.0 PERFORMANCE OF THE MODEL

We characterize the performance of the model described in Section II in terms of the expected number in the system L,

$$L = \sum_{n=1}^{K} n p_n \ ,$$

the expected system time T (waiting time plus service time --- from Little's result [5, p. 17]),

$$T = L/\lambda(1 - p_K),$$

and the rate of missed customers M,

$$M = \lambda p_K \ .$$

We study the overall behavior of the model by seeing how L, T, and M are affected by the parameters λ, K, and α or β.

3.1 Performance for Linear Service Times

We start with the case in which the expected service time increases linearly with the number of customers in the system (Eqs. (4)-(6)). For fixed system capacity K = 50 and fixed load factor $\rho = \lambda/\mu = .5$, Fig. 1 shows the expected number in the system L and the expected system time T as functions of α. Performance is relatively unaffected by increases in α until α reaches a threshold, at which point the performance suddenly collapses --- the system saturates with L approaching K and with T growing unbounded. Furthermore, as L approaches K, p_K approaches 1 so that the missed customer rate M approaches λ. If the second term in $1/\mu_n = (1/\mu) + (\alpha/\mu)(n - 1)$ is thought of as an overhead term, Fig. 1 suggests that there are regions of system operation in which a small change in the overhead "rate" α will lead to a performance collapse. In its suddenness, the phenomenon is reminiscent of thrashing [6].

The sensitive region of operation can also be encountered by changes in ρ. For example, Fig. 2 shows L and T as a function of ρ for K = 50 and $\alpha = .04$ and also for $\alpha = 0$. The $\alpha = 0$ curves show the no-overhead M/M/1/K case. The $\alpha = .04$ curves show that a small change in the arrival rate λ can result in a performance collapse.

The performance of M/M/1/K systems can always be improved by increasing the storage capacity --- in the $K \to \infty$ limit (M/M/1), L and T reach limits and M goes to zero. In sharp contrast, the performance of the modified model can collapse when K is increased. Figure 3 shows L a a function of K for three values of α --- $\alpha = 0$ (M/M/1/K), $\alpha = .03$, and $\alpha = .06$ (the load factor is $\rho = .5$ in all cases). As K increases, the curves for $\alpha = .03$ and $\alpha = .06$ eventually increase as fast as K. The effect can be seen in different terms

in Fig. 4, which shows p_K vs. K. For the M/M/1/K case ($\alpha = 0$), p_K decreases monotonically and is asymptotic to zero. But when $\alpha \neq 0$, p_K (and therefore M) eventually increases with K, owing to the possibility of large service times, and in all cases p_k is asymptotic to one instead of zero! These somewhat surprising results suggest that there are situations in which adding memory to a system will excacerbate rather than solve a performance problem.

3.2 Performance for Logarithmic Service Times

One might suspect that the performance collapse would be much less dramatic if the increase in expected service time $1/\mu_n$ were much less than linear. However, similar effects occur for logarithmic increases (Eqs. (7)-(9)), although for considerably larger values of the proportionality constant β than was the case for α. For example, Fig. 5 shows L and T as functions of β for system capacity K = 50 and load factor $\rho = .5$, and Fig. 6 shows L and T as functions of ρ for K = 50 and $\beta = .3$. By comparing Figs. 2 and 6, one can see that performance in the logarithmic case can change almost as much for a small change in ρ as can performance in the linear case. Finally, Fig. 7 shows L as a function of K for three values of β --- $\beta = 0$ (M/M/1/K), $\beta = .3$, and $\beta = .4$ (the load factor is $\rho = .5$ in all cases). Although the curves are not as steep as those in Fig. 3, the effect is still dramatic.

4.0 DYNAMIC BEHAVIOR

Equilibrium results such as those in Figs. 1-7 are interesting and

informative, but they shed little light on the dynamic behavior of a system. For example, Fig. 1 predicts that a system with K = 50, ρ = .5, and α = .06 will operate with an average queue length of about 49, but it gives no information about how fast an initially-empty system will deteriorate. Indeed, equilibrium results like Fig. 1 can be misleading, since it is easy to get in the habit of assuming that a short-term time average of the queue length will increase at a steady rate until the equilibrium expected value is reached and that the system will thenceforth operate with queue lengths close to the equilibrium expected value. The performance collapse model is a dramatic counterexample to such thinking.

A first hint at the system's dynamic behavior can be obtained by studying Fig. 3. The shape of the curves suggest that an initially empty system might operate with good performance for quite some time and then degrade suddenly after statistical fluctuations have caused the queue length to exceed some threshold value. More precisely, when the system is in a state with queue length $n \neq 0$, the relatively likelihood of a departure vs. an arrival is given by the ratio $\rho_n = \lambda/\mu_n$: A departure is more likely than an arrival when $\rho_n < 1$ and vice versa. Since μ_n decreases monotonically as n increases (see (1)), there will be a bounded region $n = \{1,\ldots,n_b\}$ in which $\rho_n < 1$ and there is a tendency for the queue length to decrease, just as there will be a region $n = \{n_b,\ldots,K\}$ in which $\rho_n > 1$ and there is a tendency for the queue length to increase. Thus, in an initially empty system there will be a kind of "pressure" that keeps the queue length down until statistical fluctuations eventually cause it to reach n_b, at which point the pressure reverses and tends to drive the queue length up. Later, statistical fluctuations should cause the queue length to drop low enough for the pressure to reverse again,

etc. One can see that the system will tend to operate in one of two quasi-stable modes --- a "low mode" with small values of n and a "high mode" with large values of n. The threshold between queue lengths in the two modes occurs at $\rho_n = \lambda/\mu_n = 1$, namely at

$$n_t = 1 + \frac{1}{\alpha}\left(\frac{1}{\rho} - 1\right) \tag{10}$$

for the linear overhead case (see (4)). Thus, as α increases, the threshold decreases monotonically and is asymptotic to 1. For small values of α or ρ, n_t may satisfy $n_t > K$, in which case there is no high mode. In the terminology of [2], n_t is a stochastically unstable congestion of locally minimum probability (p_n has a local minimum in the vicinity of $n = n_t$).

I confirmed the foregoing by studying the dynamics of a variety of test cases by means of simulation. Here are the results of one such run, conducted for the case of linear overhead with K = 40, $\rho = .5$, and $\alpha = .06$ (Fig. 3 predicts an equilibrium average queue length of about 30): After starting at n = 0, the queue length didn't exceed n = 9 during the first 4100 transitions (arrivals or departures). The average queue length during this period was 1.52. The queue length then went up and hovered in the vicinity of n = 17 for about 320 transitions (for this case the mode threshold is $n_t = 17.66$), after which it went up quickly, hovered in the vicinity of n = 38 for about the next 14,000 transitions, and then came down again to the vicinity of n = 0. The average queue length during the high mode was 38.15 and the shortest queue length was 23. The overall queue length average was 29.9. In general, for fixed values of K and ρ, increasing α results in a greater portion of time spent in the collapsed state.

One can study this more formally in terms of first passage times in the embedded Markov chain on the integers $n = \{0,1,\ldots,K\}$. The transition

probabilities $p_{i,j}$ from states n = i to states n = j are

$$
\begin{aligned}
p_{0,1} &= 1 , \\
p_{j,j-1} &= \frac{\mu_j}{\lambda + \mu_j} \quad , \quad 1 \leq j \leq K-1 \\
p_{j,j+1} &= \frac{\lambda}{\lambda + \mu_j} \quad , \quad 1 \leq j \leq K-1 \\
p_{K,K-1} &= 1 , \\
p_{i,j} &= 0 \quad , \quad i \neq j \pm 1 \quad .
\end{aligned} \tag{11}
$$

Let $m_{i,j}$ be the expected number of transitions from a time the system is in state n = i until it first reaches state n = j. Since the embedded Markov chain is irreducible and recurrent, the $m_{i,j}$ are given [7, p. 243] by the system of equations

$$
m_{i,j} = 1 + \sum_{k \neq j} p_{i,k} \, m_{k,j} \quad . \tag{12}
$$

For the transition probabilities given by (11), the solution of (12) is [8]

$$
m_{i,j} = \sum_{h=1}^{j-1} \frac{1 + \sum_{k=1}^{h} (1 + \rho_k) P_{k-1}}{P_h} \tag{13}
$$

for $0 \leq i < j \leq K$, where $P_n = \rho_1 \rho_2 \cdots \rho_n$. To obtain m_{ij} for $i > j$, one writes $1/\rho_{L-n}$ for ρ_n in (13) and P_n, and then computes $m_{L-i,L-j}$. For example, given the parameter values used in the simulation mentioned above (K = 40, ρ = .5, and α = .06), the expected values $m_{1,17}$ and $m_{40,17}$ are

6354 and 15026 respectively. The observed values in the (single run) simulation were about 4100 and 14000 respectively.

Since a change from one of the quasi stable modes to the other is followed by a relatively long period of operation in the new mode, we take $m_{0,K}$ and $m_{K,0}$ as approximate measures of the relative amounts of time spent in the low mode and high mode respectively, and we define

$$F_h = \frac{m_{K,0}}{m_{K,0} + m_{0,K}} \tag{14}$$

as an approximate measure of the fraction of time spent in the high mode. The mode fraction F_h and the transition point n_t (Eq. (10)) provide information about the model's dynamic behavior, and can be used to supplement equilibrium results such as those in Figs. 1-7. As an example, for linear overhead with parameters K = 50 and $\rho = .5$, we plot F_h and n_t as functions of α in Fig. 8 (compare Fig. 1). As α increases, the mode threshold decreases and the system spends an increasing fraction of time in the high mode.

It is important to realize that the system can continue to spend long periods in the low mode even when F_h is close to 1. For example, from Fig. 8 we see that $F_h = .99$ when $\alpha = .056$, but the expected first passage time is still large at $m_{0,50} = 22{,}245$ transitions (more than 10,000 customers). Of course $m_{50,0}$ is about 100 times longer, but the large value of $m_{0,50}$ suggests the interesting possibility of a dynamic operating procedure that leads to overall good performance: Values of F_h close to one correspond to equilibrium situations in which there is a large probability of a full system. (For example, Fig. 4 shows that p_{50} is almost .5 when $\alpha = .06$, K = 50, and $\rho = .5$.) In such cases, many customers will be turned away at

rates p_K for long periods of time on the order of $m_{K,0}$. It makes more sense to turn away all customers for the relatively short time it would take to serve all customers currently in the system, and then to resume accepting customers when the system is empty.

5.0 DISCUSSION

Intuition about the performance of computer systems is notoriously bad --- we always seem to underestimate the powerful combinatorial effects involved. Queuing models help because among other things they warn that seemingly innocent modifications can have disastrous effects. The model considered in this paper has this quality --- it shows how a small amount of overhead per customer can choke the system, and it shows how more memory is not always the answer to poor performance.

Furthermore, the model shows how equilibrium results can be misleading with respect to dynamic behavior, and how they can can obscure the possibility of performance gains available by means of dynamic operating procedures.

6.0 ACKNOWLEDGEMENTS

After listening to me desribe performance collapse in a more elaborate model, Carl Landwehr suggested the simple version considered in this paper. I thank him for the suggestion, as well as for several helpful discussions. I also thank Rodney Johnson for stimulating the consideration of dynamic behavior, for finding Eq. (13), for providing programming assistance, and for discussing various aspects of the results with me.

References

1. A.O. Allen, Probability, Statistics, and Queueing Theory With Computer Applications, New York, Academic Press, 1978.

2. P.-J. Courtois and H. Vantilborgh, "Stable and Unstable Congestions in Markovian Queues," MBLE Research Rept. R382, Philips Research Laboratory, Brussels, September, 1978.

3. N. Abramson, "Packet Switching with Satellites," Proc. AFIPS 1973 NCC, AFIPS Press, Montvale, N.J., pp. 695-702.

4. L. Kleinrock and S. Lam, "Packet Switching in a Mudlti-Broadcast Channel: Performance Evaluation," IEEE Trans. on Communications COM-23, (April, 1975), pp. 410-423.

5. L. Kleinrock, Queueing Systems, Vol. I, New York, John Wiley, 1975.

6. P.J. Denning, "Thrashing, Its Causes and Prevention," AFIPS Conference Proceedings (1968 FJCC), pp. 915-922.

7. E. Parzen, Stochastic Processes, San Francisco, Holden-Day, 1962.

8. R.W. Johnson, private communication.

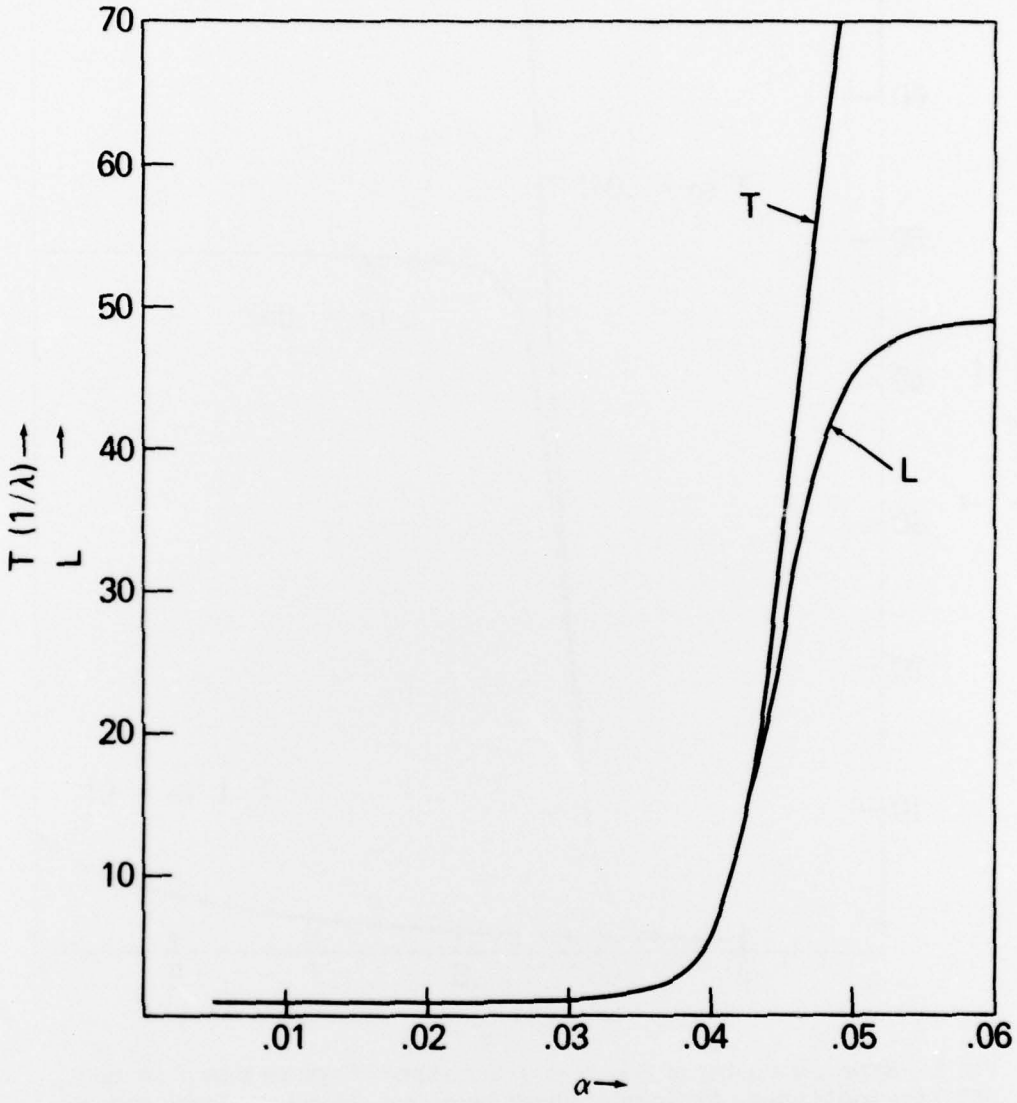


Fig. 1 — Expected number of customers L and expected system time T (in units of $1/\lambda$) vs. linear overhead factor a for system capacity K = 50 and load factor ρ = .5.

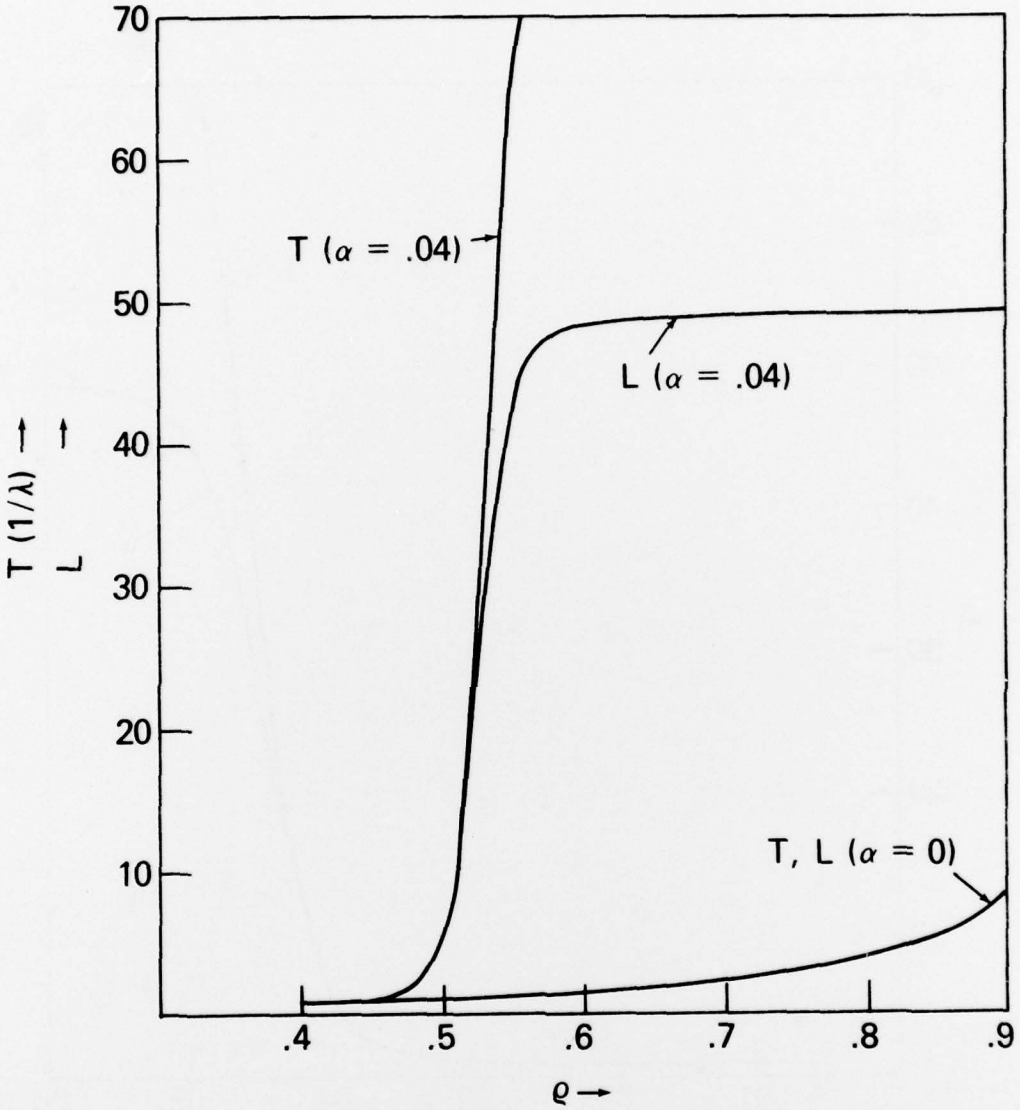


Fig. 2 — Expected number of customers L and expected system time T (in units of $1/\lambda$) vs. load factor ρ for linear overhead factors $a = .04$ and $a = 0$ with system capacity K = 50. The $a = 0$ curves show performance for the overhead-free M/M/1/K case.

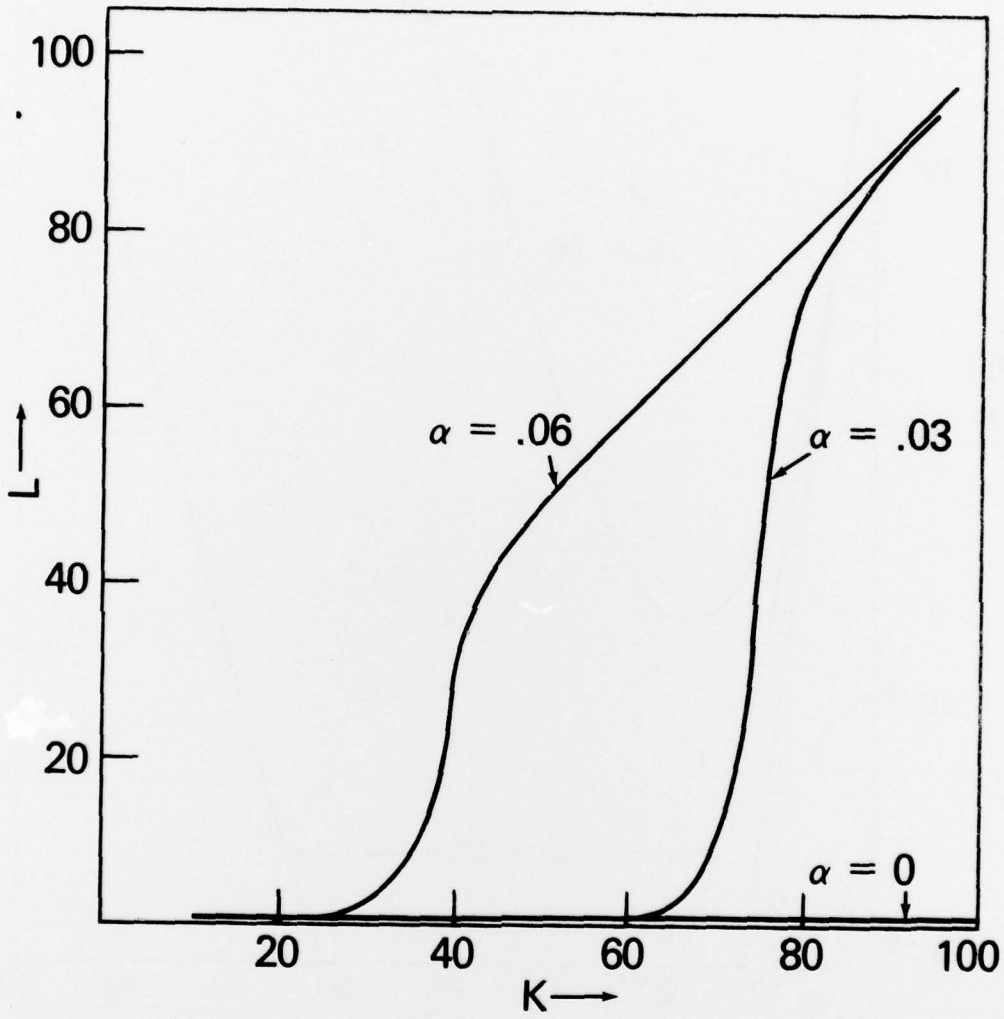


Fig. 3 — Expected number of customers L vs. system capacity K for three values of the linear overhead factor α. The load factor is $\rho = .5$.

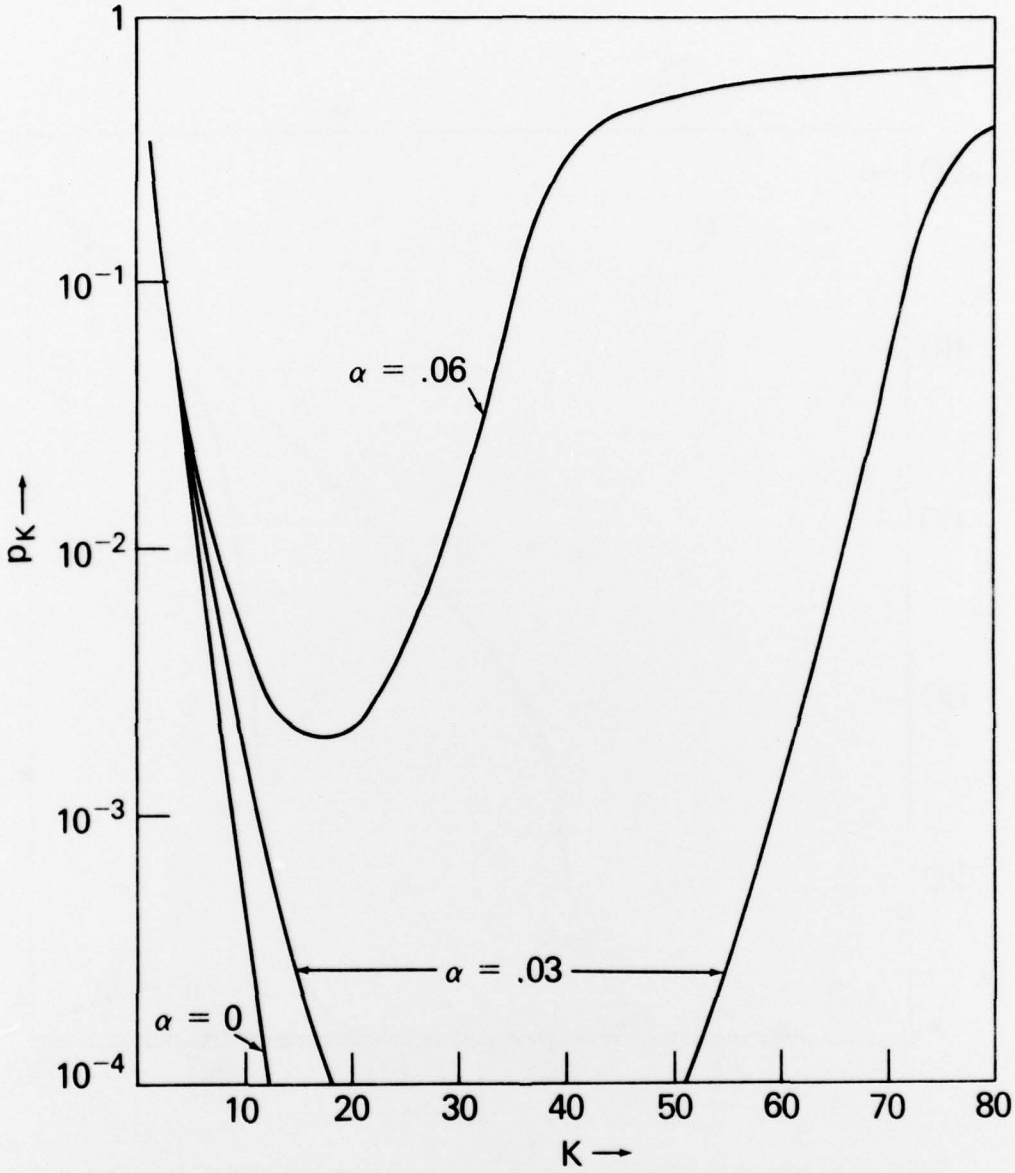


Fig. 4 — Probability that the system is full vs. system capacity K for three values of the linear overhead factor a. The load factor is $\rho = .5$.

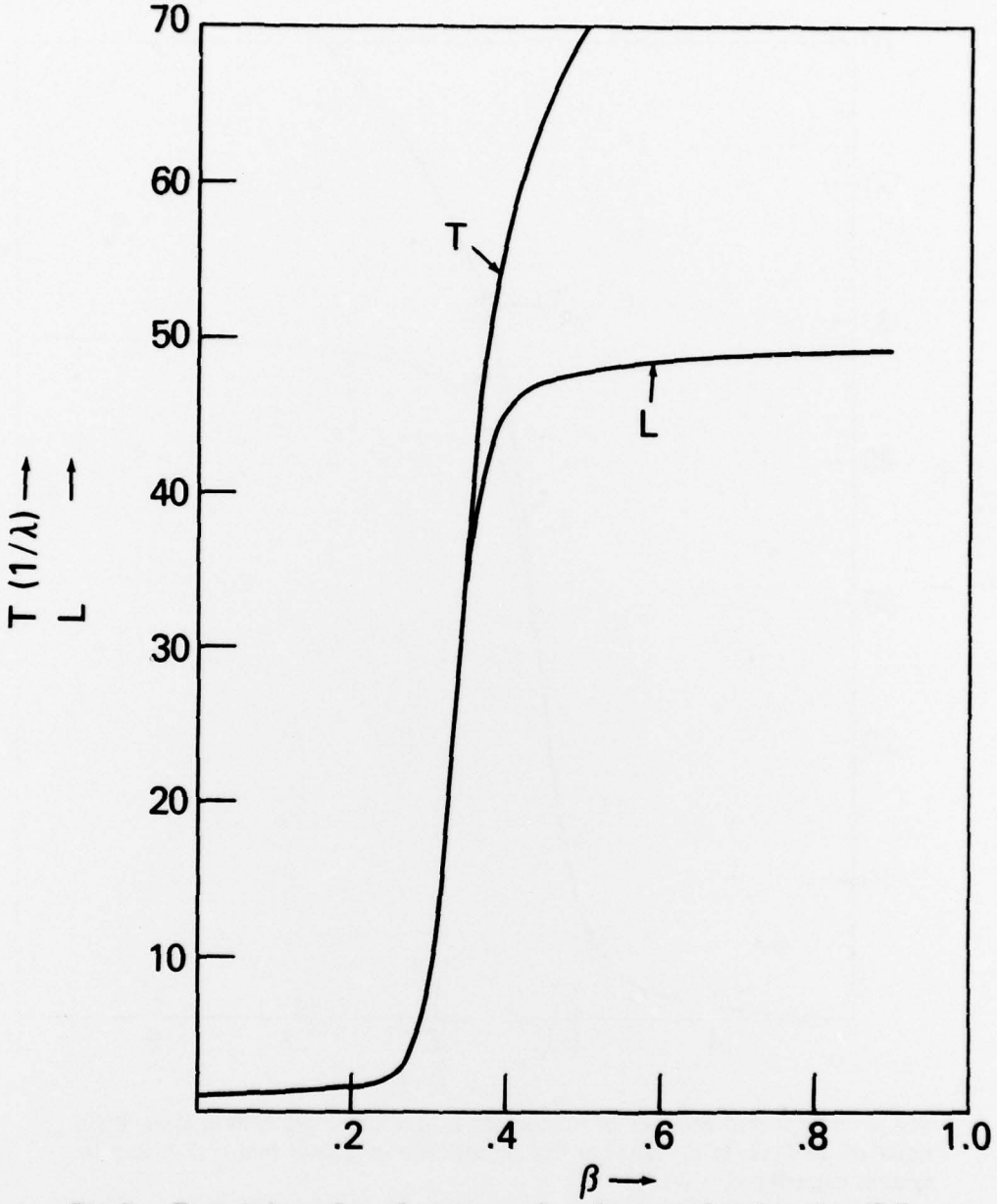


Fig. 5 — Expected number of customers L and expected system time T (in units of $1/\lambda$) vs. logarithmic overhead factor β for system capacity K = 50 and load factor ρ = .5.

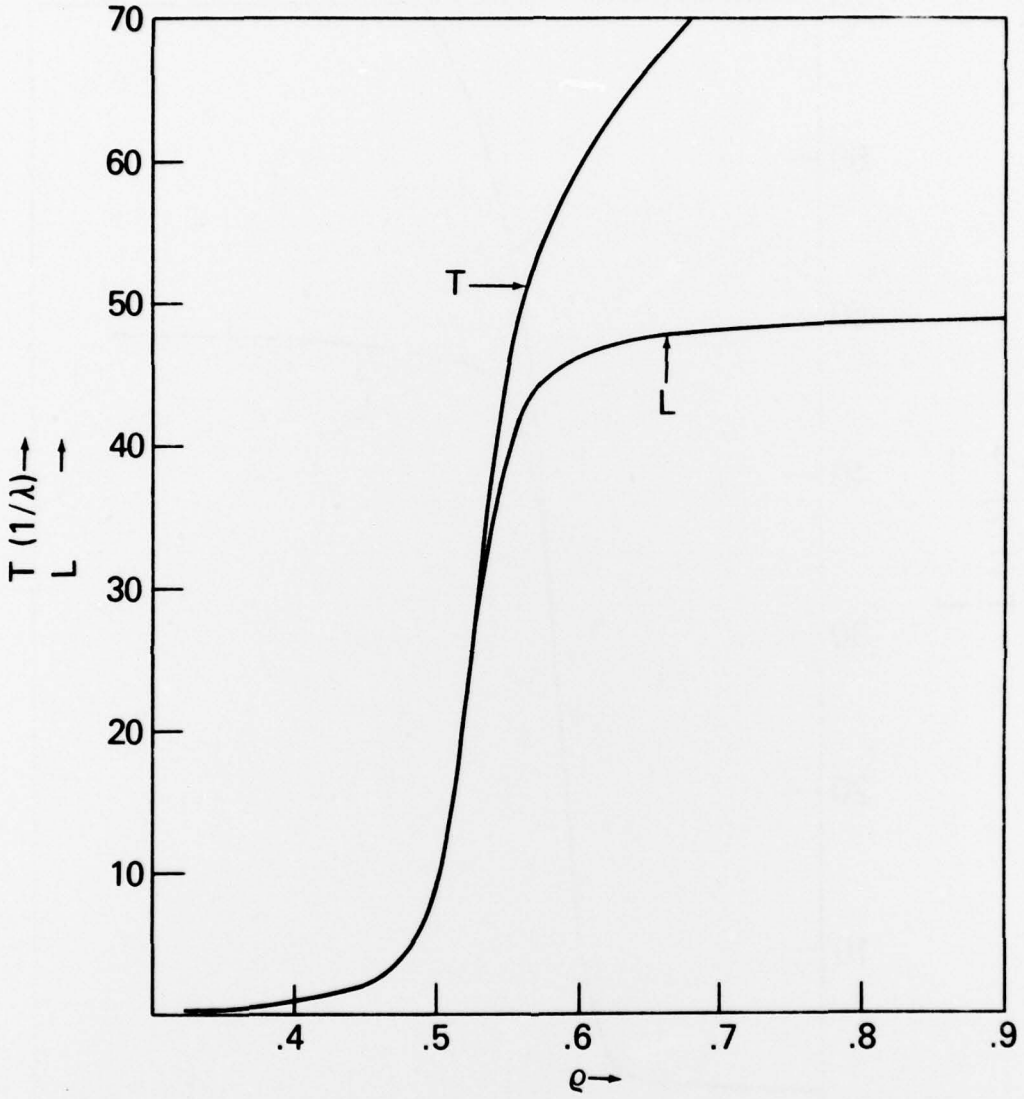


Fig. 6 — Expected number of customers L and expected system time T (in units of $1/\lambda$) vs. load factor ρ for logarithmic overhead factor $\beta = .3$ with system capacity K = 50.

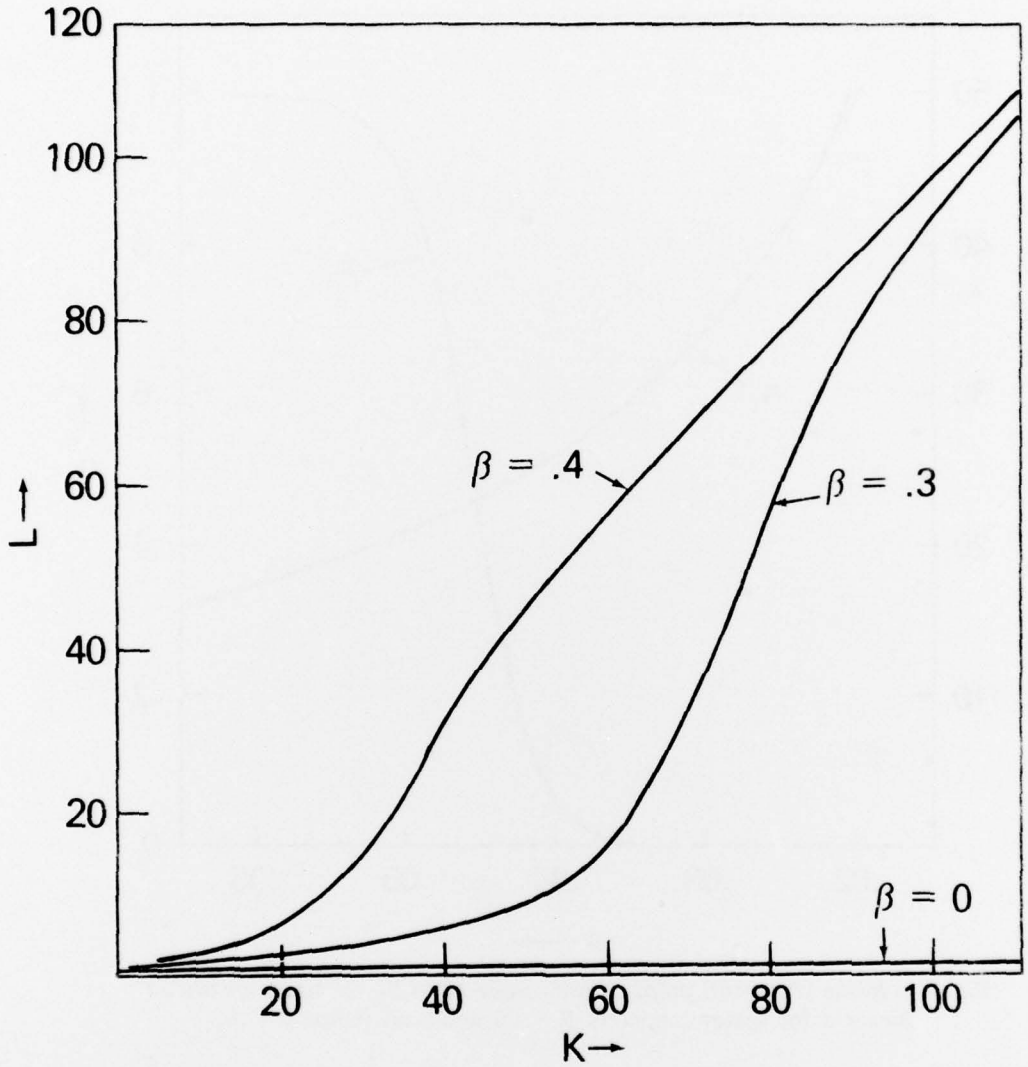


Fig. 7 — Expected number of customers L vs. system capacity K for three values of the logarithmic overhead factor β. The load factor is $\rho = .5$.

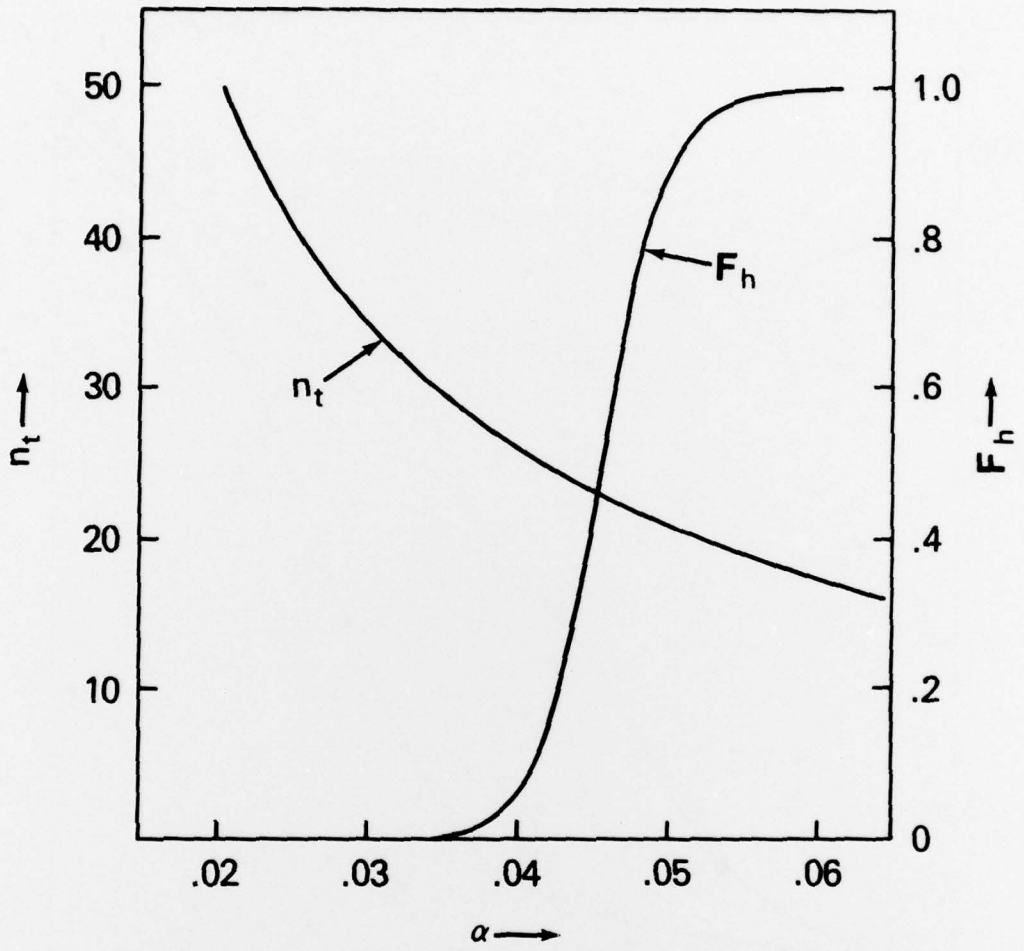


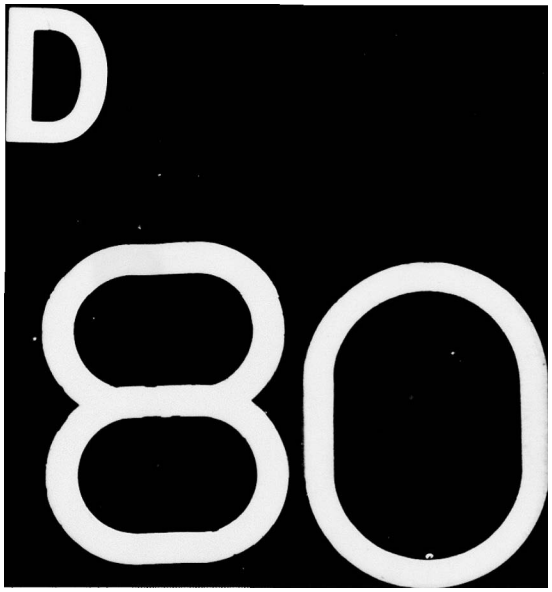
Fig. 8 — Mode transition point n_t and mode ratio F_h vs. linear overhead factor α for system capacity K = 50 and load factor ρ = .5.

DEPARTMENT OF THE NAVY

NAVAL RESEARCH LABORATORY
Washington, D.C. 20375

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

POSTAGE AND FEES PAID
DEPARTMENT OF THE NAVY
DoD-316
THIRD CLASS MAIL


D
80